23 August 1983

MEMORANDUM FOR: The Record

FROM: _____

OP/HRPS

SUBJECT:

2. Although the presentation emphasized methodology and mathematical model building in the employment discrimination area, no references to Agency cases were cited nor did we receive any case specific questions. Questions and comments from the panel discussants Prof. Scott (Univ. of California – Berkeley) and Dr. Taeuber (Smithsonion) centered on the importance of biographic vs. merit variables in model building, the empirical statistical distributions for salary derived from the regression equations, and a renewed need for data validity and model reliability. Floor questions also centered on biographic vs. merit variables in regression models and the audience seemed pleasantly surprised that our R^2 values were as high as we reported, e.g., 85-90 percent. R^2 measures the appropriateness and accuracy of models with 100 percent the theoretical ideal. Prof. Scott briefly reported out preliminary R^2 values from her on going studies in the 70-72 percent range and the people from Rutgers University and Bell Labs even smaller. The differences between their values and ours should provide the Agency with an appreciation that Agency models are equal to, if not better than, that obtained elsewhere according to the R^2 criterion.

3. Approximately 125 were in attendance at our session. This is a gratifying number considering we got underway at 8:30 A.M. on Monday. The floor session ran over the allotted time and we continued with an informal session outside the Grand Ballroom West of the Sheraton Center for approximately another 45 minutes. Private discussions were held with Dr. James Cole, co-author of Statistical Proof in Discrimination (with Prof. Baldus), Prof. Scott, and other attendees.

4. George and I could not ascertain any negatives with respect to the substance of the paper. This too is gratifying and we were surprised at the number of prepublication and prepresentation copies that were requested from academic, industry and government. In all, we distributed approximately 150 copies before and after the presentation. Publication will be forthcoming in the 1983 Proceedings of the American Statistical Association.

5. Did have an opportunity to meet several notable authors, researchers and practitioners from academia and the applied world and also share experiences in struggling through data to derive models to prepare for court. All in all a long trip but worthwhile.

STAT

EVALUATING THE EFFECTS OF GENDER

IN EMPLOYMENT DISCRIMINATION CASES:

JURIMETRIC DATA BASES, DATA RELIABILITY AND

STRATEGIES FOR USING REGRESSION MODELS.

by
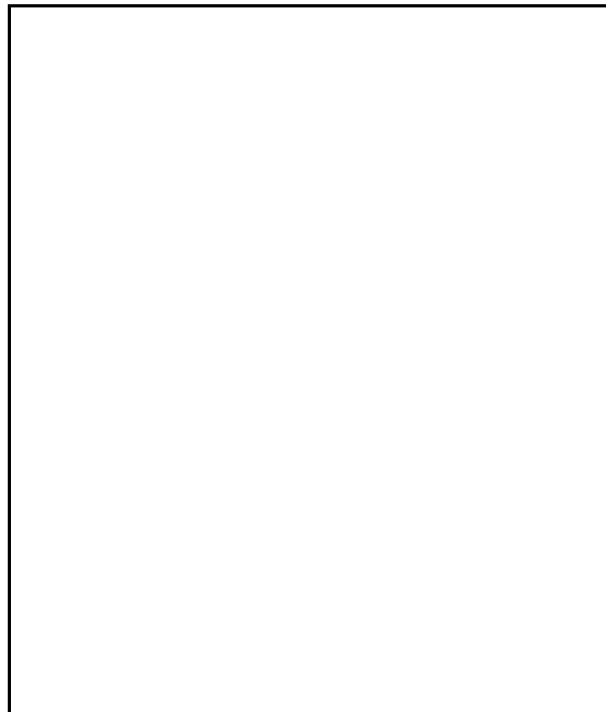
STAT

1.   Introduction

Various authors (Connelly-Peterson, Kaye, McCabe, Reich) have
pointed out that statistical methods are being used, at an
increasing rate, to study sex and/or wage discrimination problems.
Moreover, the resolution of issues in many litigation cases requires
the use of quantitative analyses or applied mathematical statistics.
Frequently, general regression models are derived and considered for
use as statistical evidence in discrimination litigation involving
class action suits and individual actions. Although model building
and variable selection procedures are crucial to such undertakings,
the foundations of such developments reside on raw or crude data
from which regression models evolve. Baldus and Cole put it quite
succinctly, "The Supreme Court has indicated that 'reliability' is
the ultimate concern in evaluating quantitative evidence" (p.3).
They define reliability in the legal context, as "the combined
properties of relevance, <u>validity</u>, accuracy, and the conceptual
simplicity that taken together largely determine the probative value
of empirical evidence" (p. 357). Moreover, "In the terminology of
the law... reliability has a much broader meaning and embraces
consistency, validity, and understandability" (p. 73).

In consequence, this paper addresses two key areas (1) data
reliability and (2) the effective use of regression models in the
legal setting. The first part of this paper addresses some of the
problems associated with statistical litigation data bases and the
reliability of litigation data, particularly the validity measure

based on incomplete or faulty data and the second part addresses in detail, model building, graphics and interpretation of results.

2.   Data Bases

We believe most organizations have large gobs of data entrenched within their personnel systems that may prove useful for regression or statistical modeling. Reviews of organizational data bases suggest that statistical litigation data bases exist in assorted forms and thus, for convenience, we tend to classify statistical litigation bases three ways, viz., anticipatory or preplanned, ad hoc and nondescript.

Anticipatory or preplanned data bases are established before a discrimination suit is filed. In short, they have been established as part of a permanent on-going activity by organizations who wish to examine, quantitatively, the potential of discrimination problems.

The advantages are many. From the viewpoint of the statistician and lawyer, there exists sufficient time to permit a thorough search for relevant data, perform research, development and testing of candidate models across a range of litigation threats, e.g., hiring, promotion, age, sex, race, etc. In our experience, the variety of court models needed is not easy to obtain. Moreover, in the absence of a suit, litigation data bases permit wide ranging statistical analyses for lawyers and managers to identify specific areas of concern. Lawyers are more than appreciative of knowing an

organization's status with respect to Title VII beforehand and
management should be aware of Title VII progress or maintenance of
compliance.

Conversely, an ad hoc data base is reactive to a pending
litigation threat and/or suit. That is, an organization finds
itself confronted with a discrimination suit and responds by
mustering whatever resources are available. This desultory
procedure presents interesting problems for statisticians in data
base construction and subsequent modeling.

One notable problem centers on the high degree of uncertainty
associated with jurists' intentions and procedures, i.e., pre-trial
and trial dates, changing legal strategies, pressures for
settlement, and the requirement to provide worthwhile statistical
analyses that must stand up in court and be completed under
sometimes frenetic conditions. Jurists readily petition the courts
for additional time along legal lines but often exhibit stout
impassivity if the statistical group needs additional time to search
for data, prospect for candidate variables and develop regression
models in a non-structured or weak data environment. A second
problem is with respect to data itself. In many instances these
data are a hodge-podge collection, insofar as mathematical
statistics is concerned, as data have evolved to meet the needs of
nonstatistical litigation requirements. In short, data are
management-oriented. In management's defense, existing data
structures were designed to meet the needs of day-to-day business
with little consideration given to formulating data bases to meet
Title VII requirements. Moreover, in those instances where a

reformulation was considered, limited resources and competing
priorities prevent implementation of specialized data bases suitable
for statistical analysis across a spectrum of litigation threats.
In this circumstance one might ask the rhetorical question, what is
the statistician to do? Most likely and perhaps none too comforting
will be the answer, do the best you can.

A nondescript data base is virtually self-defined. It is,
however, instructive to point out to jurists that in searching for
candidate variables and data deemed important it is not too uncommon
to find a multiplicity of data voids, e.g., archival records and
history tapes only go back so far; data on certain variables were
deleted or changed a few years ago and in some cases data are
nonextistent. We all know too well that multiple voids in data sets
provide discouraging R^2 values when data sets are fragmented.

2.1   Legal Framework for Data Bases

Our experience with defense counsel indicates that they wish to
establish three things with respect to data bases: (1) current
status, with the most recent data (analysis) that is available (2)
provide data (analysis) to the closest date when the suit was filed
and (3) thereby present a dynamic picture of Title VII compliance,
improvement or degradation. In this two state or multi-state data
base procedure, data are often heterogeneous and thus occasionally
complicating as some variables may not have the same meaning in two
or more time frames.

Frequently additional data digging occurs with no guarantee of success in finding suitable and non-tainted data (variables) that will yield models with R^2's, residuals and AOV's that are worthwhile. In some cases it may be fortuitous, due to the legal environment (no pre-trial date has been set) and the willingness to persevere statistically, that successive and improved data bases can be generated such that the final data base becomes worthwhile for analyses. However, don't count on it. Hard work, diligence and research are no guarantee. Lawyers ought to know this at the outset and be reminded frequently to cushion their disappointment if it comes.

2.2   Benchmarks for Data Base Quality

In any data environment the statistician will go prospecting for data and candidate variables. We advocate that early in data exploration and data base construction the realization of data reliability or validity will dominate the yen to run a few regressions. The prescription for preplanning, specifying objectives and checkpoints for progress, as given in Draper and Smith (p. 418) is an excellent road map to follow.

Check the numbers carefully as soon as possible — decimal points are always being misplaced; the quality of "messy data" is usually fairly poor. Do not attempt to build a

model on a set of poor data! ... one often finds 14 inch

men, 1,000 pound women, students with "no" lungs, and so

on. All the planning and training in the world will not

eliminate these sorts of problems. These are "human"

errors, not computer errors, although the computer often

gets the blame. In our decades of experience with 'messy

data", we have yet to find a large data set completely

free of such quality problems.

In our experience we have found predictor-variables such as

length of service (tenure) to exceed 200 years, college degree

conferred before birth, salaries several orders of magnitude greater

than received, and so on. Plots of basic variables against the

dependent variable also helps to identify these maverick

observations. Some computer programs provide univariate statistics

whereby maximum and minimum values for observations are identified.

A rule of thumb we have applied is to check everything with whatever

devices are available and then replot basic variables against the

dependent variable - again. Time spent checking a data base is as

crucial as finding the best model and, frankly, we believe it is

more important.

A similar issue is to validate or verify the special data bases

constructed for statistical analysis. Frequently, statisticians

resort to a two-stage process whereby crude data are tapped from a

variety of sources, i.e., machine and manual, and then entered into

a special data base for statistical analysis. This straight forward

procedure can also be a source of error. Although reliable data

from other machine sources is less suspect, re-entry of data derived from manual sources has the usual opportunities for error, particularly in the subjective interpretation of numbers.

Consequently, the straight-forward procedures of acceptance sampling has the utmost utility. One procedure found useful is to take two or three small samples from each data base, both crude and special, for each time frame. Such a procedure can provide interesting results. Seemingly, one or two data fields, or variables, exhibit high error rates, while others may vary around the so-called nominal rate of five percent and others are virtually error-free. The choice of an acceptable error rate is an interesting question in its own right. Arbitrarily we use and suggest five percent or less as the target number and believe this value to be acceptable in the courts. Obviously, data fields in excess of this nominal value should be rectified.

Missing values also pose interesting problems. These occur from a variety of causes, e.g., individuals performing data entry may not have values to enter and may conveniently enter missing value codes and ignore such values when updating. In certain instances this may not be a serious problem but it can become serious if many updates on many fields are performed over long periods of time. In dealing with a large litigation data base, e.g., a divisional population, it is a dolorous situation when a significant portion of population values are missing. The importance and existence of an on-going quality control system is paramount to minimize the impact of messy data.

If one utilizes what we call the Scott methodology, i.e.,
derive the best male prediction equation as the basis to estimate
female salary, then much attention should be given to homogeneity or
equivalence of groups. That is, there may be categories of males
inwhich there is no female counterpart and vice versa. An
organization may have a large number of male engineers and no female
engineers. To include these data in the analysis is misleading.
Cell counts of both male and female categories are required with
appropriate deletions of all categories (male and female) in which
there is no one-to-one category or cell correspondence between males
and females. The task is to establish a reduced data set to meet
Scott's homogeneity or equivalence criterion.

2.3   Recommendations Concerning the Importance of Data Reliability

As defined and explained by Baldus and Cole (p.273), the common
threats to validity in regression analysis, in the discrimination
context, center on: (1) problems with variable selection (too many
vs. too few, multicolinearity, merit vs. biographic [cf. McCabe]),
(2) mathematical structural assumptions (linearity, additivity,
interactions), (3) sampling error (non-experimental design), and (4)
error term behavior (residual distributions, validity of statistical
tests). These four points are crucial when verifying or challenging
the assertion that regression methods rest on a firm mathematical
foundation. We urge that data reliability be added to the list.
Our experience indicates that the first four are usually given

considerable treatment, however, if they are based on data assumed
reliable (the vanity effect) then the best of analysis is suspect.
Moreover such analysis, when based on faulty data, seriously
undermines the need to directly support the burden of proof and
indirectly support the burden of persuasion.

3.   Strategies For Using Regression Models

As cited in the introduction, the purpose of statistics in
legal proceedings is to quantify evidence from historical records or
information about the defendent and/or plaintiff. There are many
methods to summarize such information ranging from simple
descriptive statistics to complex regression models. This section
concentrates on the effective use of regression models in the legal
setting and on modes for presenting the results of such complex
analyses to jurists. In particular, regression models are used to
predict salary as a function of the persons qualifications in an
attempt to determine whether there is inequity between the salaries
of professional males and females. The hypothesis is that males and
females with the same qualifications receive equal pay. The use of
regression analysis is to determine if there is evidence for or
against the hypothesis.

The next section of this paper presents two methods to build
models to describe salary as a function of qualifications, which is
a combination of ideas from Scott (1979) and McCabe (1979). The two
methods consist of building 1) a Best Male Model (or opposite sex

model) and 2) a two sex model. Also, the next section contains graphical and statistical methods to help interpret the findings of the model building process. Section 3.2 shows how to make adjustments in salary when inequities are found. And finally, Section 3.3 discusses how management can use the regression models to monitor their salary program and, if necessary, make adjustments to correct for actions which have led to the salary inequities.

3.1   Strategies for Selecting Models Which Are
Adequate and Appropriate For Use in Legal Procedings

Effective modeling depends on the quality of the data base, thus before modeling can be considered, the statistician must certify the accuracy of the information in the data base. Most data bases contain many more variables about an employee than are needed to describe salary. The first cut on the types of variables in the data set is to determine which variables are appropriate to be used in a model to describe salary. For example, information about an employee's hobbies is most likely not appropriate to describe salary. The second cut on the types of variables is to select those which are appropriate for use in models for legal proceedings involving gender discrimination. Those predictor variables which are appropriate for use in discrimination models are called biographical variables. Examples of biographical variables are 1) number of years on current job, 2) number of years on a similar job before the current job, 3) degree, 4) years since highest degree, 5)

grade or level at time of hire, 6) other skills and/or skill levels which have been acquired such as additional degrees in related fields or experience through short courses or other training programs, and 7) etc.. The key for the biographical variables is that they are easy to measure and their value and accuracy are seldom subject to dispute (McCabe, p.27).

Those predictor variables which are not appropriate for use in models for discrimination proceedings are merit variables such as 1) production as measured by output on the current job (e.g., through the number of important papers published, number of important committees, or number of important clients or accounts), 2) current rank and 3) ratings by supervisors. The merit variables may be difficult to measure and their accuracy may be disputed (e.g., which clients are important, etc.). Measurement of merit variables could be tainted or biased as their values could be the result of discrimination. For example, if the supervisor discriminates against one sex, a person of that sex may not receive as high a rating as one should, or they may not be assigned as good of clients, accounts, or committees as are assigned to their favored sex counterparts. Models with tainted variables should not be presented in court. When models with tainted variables are used in court, it is up to opposing counsel to identify such variables as tainted and argue that information from such regression models must not be allowed as evidence in the case (see Finkelstein (1979) for examples). Also, one must be careful that chosen biographical variables) are not tainted. For example, discrimination may occur in the process of selecting the people to attend a short course to

upgrade a certain skill. Thus, one must justify and document that such training variables are not influenced by discrimination.

Once the appropriate predictor variables have been selected, they can be used to build a model to describe salary (the model is to describe the mean salary of the people given the values of the predictor variables). We do not know the correct model to determine the mean salary as a function of the predictor variables, so we set out to construct a model which is adequate. An adequate model is a model which does a very good job of describing the mean salary as a function of the predictor variables in the range of the predictor variables. One measure of adequacy is the value of R^2. We like a model to have a moderately large R^2 value, i.e., in the range of .8 to .9, though some models have been used with R^2's in the range of .5 to .7. Another measure of model adequacy is to test for lack of fit whenever possible. But in any case, the residuals should always be examined for possible patterns.

When building a regression model, there are two possible objectives, depending on the group constructing the model. The statistician for the defendent is trying to use the regression model to show there is no evidence to substantiate discrimination. The statistician for the plaintiff is trying to use the regression model to show there is evidence for discrimination. Each side should build their models from the biographical variables and each should document the criterion and paths used to build their respective models. Both models should be adequate to describe the mean salary, though the models may contain somewhat different variables depending on the model building method used (i.e., backward or forward or

minimum mean square residual or a method using the Cp statistic, etc.)

The models can contain independent variables which are quantitative as well as independent variables which are qualitative (classification type of variable). When using classification variables in a model building procedure, we have been successful in carrying out a backward elimination procedure with SAS[TM1] PROC GLM. The GLM program allows the use of both class variables and continuous variables. This permits a model to be constructed which has several intercepts (for each class) and several slopes (for each quantitative variable and interaction of quantitative variables with eachother and with the class variables).

There are two strategies for building salary models for use in the discrimination context (both are referred to in Scott (1979)). The first approach is called the BEST MALE MODEL and the second approach is called the TWO SEX MODEL, each approach is described.

-[1]SAS is a registered trademark of SAS INSTITUTE INC.

BEST MALE MODEL

The BEST MALE MODEL approach consists of constructing an
adequate and appropriate model to describe the male salaries and
then use the same model to predict the female salaries. The task is
to compare the differences between the actual female salaries to the
predicted female salaries (from the best male model), called DSAL's.
We want to see how well the best male model predicts the actual
female salaries. There are many ways to plot the DSAL's from
various types of frequency distributions (as used by Scott (1979))
to multi-way plots. We have found it informative to plot the DSAL's
vs the independent variables in the models, often for various
combinations of the class variables (e.g., plot vs time on the job
for each entry level).

Figures 1, 2 and 3 are plots of DSAL's vs time of employment
where Figure 1 shows no evidence of discrimination, Figure 2 shows
evidence of discrimination where females are hired with pay equal to
their male counterparts but their rate of increase is slower, and
Figure 3 shows evidence of discrimination where females are overpaid
in the early years in an attempt to attract females to the job, but
the females still progress slower than the corresponding males.
Figure 3 shows two types of discrimination. The first type is when
females are paid more than equally qualified males earlier in their
careers. The second type is when the rate of increase in pay is
slower for females than for males.

The statistician for the defense hopes for DSAL plots like
Figure 1 while the statistician for the plaintiff hopes for DSAL

plots like Figure 2. The DSAL plot in Figure 3 presents problems
for both sides.

The DSAL's can be combined for various ranges of the
independent variable to provide a simpler picture. Figures 4, 5 and
6 are plots of the mean and range of the DSAL's for five year
intervals and are much cleaner than those with the individual
DSAL's.

If there are patterns in the DSAL plots, the analysis could be
continued by including the possibly tainted merit variables in the
model to see if a more adequate model can be obtained. The result
would be to provide a possible answer to the source of the
discrimination.

The best male model can also be used to investigate
discrimination against an individual. The process is to fit the
best male model and compute the male residuals from the model
(denoted by RES) and compute the female DSAL's. Plot the RES and
the DSAL's vs the independent variables and the identity of the
plaintiff on the DSAL plot. Overlay the plot with the plaintiff's
DSAL on the corresponding male RES plot. Then determine if the
plaintiff is worse off than some of her male counterparts. Figures
7 and 8 show there is possible discrimination, but it is not sexual
discrimination as there are both males and females with much lower
salaries than expected. Figures 9 and 10 show that the plaintiff as
well as some other females are bei ng discriminated against, though
these does not seem to be discrimination against the class as a
whole.

There are two disadvantages of the best male model approach.
First there are no formal testing procedures to provide significance
tests for the no discrimination hypothesis. Second, the values of
some of the biographical variables may be tainted because of
discrimination in society rather than by the defendant. For
example, a female may not have had the opportunity (due to society
biases) to pursue the desired advanced degree prior to her
employment, thus putting her at a disadvantage at the start.

TWO SEX MODEL

The two sex model strategy uses all the data to build the
regression model, i.e., data from both sexes are used to construct
one model. The two sex regression model must be built by allowing
for a different slope for each sex for each independent variable and
a different intercept for each sex for each class variable. For
example, if there are three continuous independent variables and one
class variable, the model that must be fit is

$$y_{ijk} = \beta_{0ki} + \beta_{1i}x_{ij} + \beta_{2i}z_{ij} + \beta_{3i}w_{ij} + \varepsilon_{ij}$$

for i=1(male), 2(female) j=1,...,n and k=1,....,p where β_{0ki}
denotes the intercept for sex i at class level k and β_{qi} denotes the
slope for sex i in the direction of the qth independent variable.
Once an adequate and appropriate two sex model has been constructed,
one can do a test of significance for equal slopes in the direction

of each independent variable. We can also test the equality of the
two regression models by using the model comparison method or extra
sum of squares method (Draper and Smith (1982)). If we fail to
reject all such hypotheses, we conclude there is no evidence for
discrimination. When two (or more) predictor variables are highly
correlated, the test for equal slopes for one variable, given
unequal slopes for the other variables, may not be rejected when in
fact it should be rejected. The model comparison method test helps
avoid that problem by testing the joint hypothesis. When the two
models are not equal, one way to investigate the source of
discrimination is to plot the predicted salaries for both males and
females on the same graph. Figures 11, 12, and 13 show typical
plots when discrimination is present. Figure 14 shows graphs where
there is no evidence of discrimination.

The advantage of the two sex model is that tests of hypotheses
can be made to compare slopes in the direction of each independent
variable as well as comparing the models as a whole. The
disadvantages are that highly correlated independent variables can
hide inequities and one large regression model must be fit to the
data (twice as large as the best male model approach).

3.2   Equitable Salary Adjustment

Several methods have been proposed to adjust salaries after the
evidence shows that discrimination has occurred. One method is to
increase the salary of each female by the same amount so that the
average female salary is equal to the average male salary. Another

method is to adjust each female's salary to the level predicted by the best male model. Both methods provide inequitable adjustments. By adding a constant to each female's salary, those who have worked one year receives the same increase as those who have worked twenty years. If discrimination has been in force for twenty or more years, a larger increase should go to a person who has worked for the company the longer period of time. The second method adjusts salaries to the mean of the male salaries with equal qualifications. Thus all of the females would have a salary less than one half and greater than the other one half of their male counterparts. The method does not allow for individual differences between females with the same values for the biographical variables.

If we make the assumption that all females have been discriminated against and that their salary relationship to the female model corresponding to the best male model is such that those females with above average abilities receive pay above the mean (or above the estimated regression model) and those with below average abilities receive pay below the mean, then a third method of adjustment should be used.

Let the best male model be denoted by

$$Y_m = x_m \beta_m + \varepsilon$$

and the corresponding female model (based on the same independent variables as the best male model) be denoted by

$$\underline{Y}_f = \underline{x}_f \underline{\beta}_f + \underline{\varepsilon}_f.$$

Let $\underline{F}$ denote the predicted means for the female salaries based on the best male model as

$$\underline{F} = \underline{x}_f \hat{\underline{\beta}}_m.$$

let $\underline{R}$ denote the residuals of the female salaries from the female model as

$$\underline{R} = \underline{Y}_f - \underline{x}_f \hat{\underline{\beta}}_f.$$

The adjusted salaries are

$$\underline{A} = \underline{F} + \underline{R} = \underline{Y}_f + \underline{x}_f \hat{\underline{\beta}}_m - \underline{x}_f \hat{\underline{\beta}}_f,$$

or the adjustment is to add

$$\underline{x}_f (\hat{\underline{\beta}}_m - \hat{\underline{\beta}}_f)$$

to the current vector of salaries.

The main property of this adjustment is that the residuals of the female salaries about the female model are identical to the residuals about the best male model. Figure 15 is a plot of the female salaries with the best male model and the corresponding female model. The second technique would adjust the female salaries to correspond to the line denoting the best male model. Figure 16 is the plot of the adjusted female salaries obtained by adding a constant to each salary so that the mean of the female salaries

correspond to the mean of the male salaries. The plot shows that there is still extreme inequity between the female salaries as well as compared to the male salaries (which were used to determine the line for the best male model). Figure 17 is a plot of the adjusted salaries obtained by applying the third method. The adjusted salaries are distributed about the female model (see Figure 15). This method of adjustment is quite complex, but is is easily carried out (when a computer program is available, like any regression program) and provides a much more equitable adjustment for all of the members of the class than does the first two methods.

3.3 Management Can Use the Information From the Regression Models

When discrimination has been found to exist in an organization, the management is obligated to correct the situation as soon as possible. Making the salary adjustments is only the first step in the correction process. Of equal importance is that they must determine the source of the discrimination and correct the problems so that discrimination does not continue. Regression models provide the tool for such an investigation. A two sex model should be built from the biographical variables and then any merit variables which increase the adequacy of the model should also be added. Next determine those variables for which there are male/female differences between the respective slopes and intercepts. Variables where there are differences are pointing toward the possible sources of discrimination. For example, if the supervisor's rating becomes a driving variable in the two sex model with unequal slopes, then

the supervisor's rating techniques must be studied to determine why discrimination is occurring. We feel that the main source of discrimination is rooted in the one, or group, which is responsible for the employee's evaluation. It will be hard to change those evaluation attitudes, but management must see that the change is made.

Most organizations do not become involved with regression models used to study discrimination until a discrimination suit has been filed against them. But a discrimination type model should be used by management to continually monitor the evaluation process as to its effect on salary and promotion. Such steps would provide a method to determine equal pay for equal qualifications and could possibly keep the organization out of court.

4.   Conclusion

Experiences with lawyers and defense counsel strongly suggest the importance of planning, developing and verifying statistical litigation data bases. We advocate that such a data base strategy should be a sine qua non in any discrimination study. It is not sufficient to take data as a given nor assume data reliability. Verification and validation should be assiduously courted. Occasionally, jurists working under pressures of deadlines find it convenient to assume that data programmed for statistical analysis is valid. They erroneously assume validity since the same data are used by an organization for the conduct of everyday business and it

is, therefore, valid. This can be devastating for both jurists and statisticians particularly if the case goes to court. Although we have found most lawyers just as sensitive to data and facts as statisticians, it is, however, incumbent upon statisticians to (1) establish the rigor of analysis (and also provide contraindicants) in structuring and validating data bases and (2) advise their legal counterparts that such undertakings are more important than running regressions and finding a model.

Very complex regression models are often needed to model salary and it is up to the statistician to present the results in an understandable medium. That medium is a graphical representation which must also be very simple. A plot of salary or DSAL's vs. one predictor variable where a plot is constructed for each subpopulation and for fixed levels of the other predictor variables has successfully been used in presenting the results to jurists.

When discrimination is found to exist in an organization, salary adjustments must be made. An equitable adjustment is described. Finally, we highly encourage management to utilize the results of such regression analyses to keep tabs on compliance with Title VII.

FIGURE 1---FEMALE DSAL PLOTS WITH NO EVIDENCE OF DISCRIMINATION
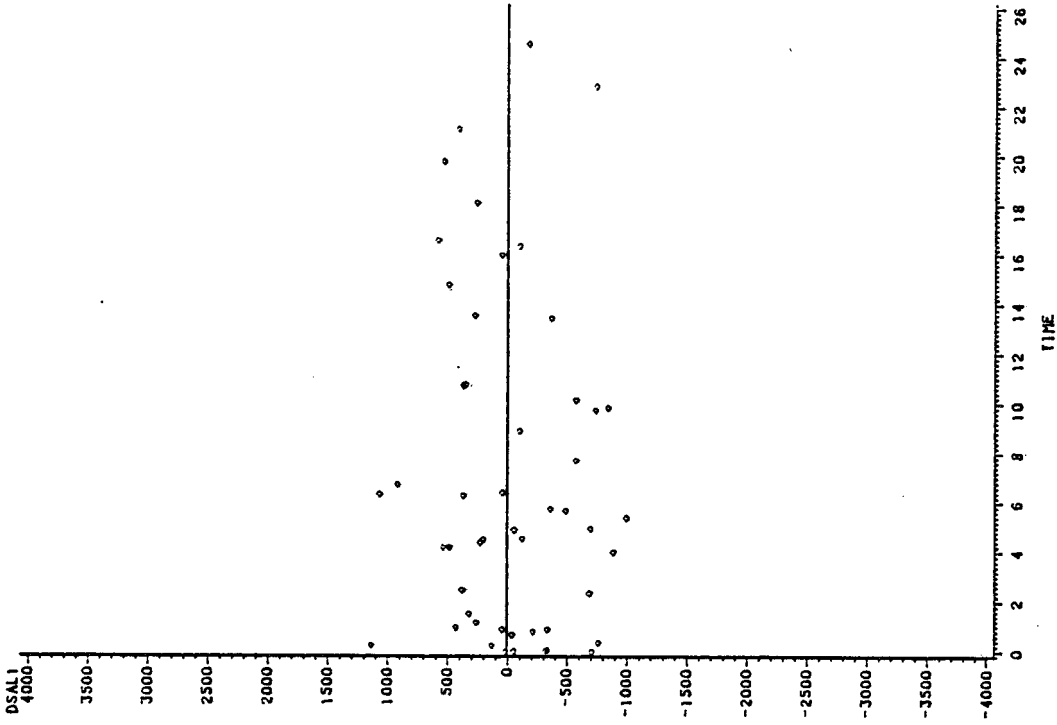
FIGURE 2---FEMALE DSAL PLOTS WITH EVIDENCE OF DISCRIMINATION
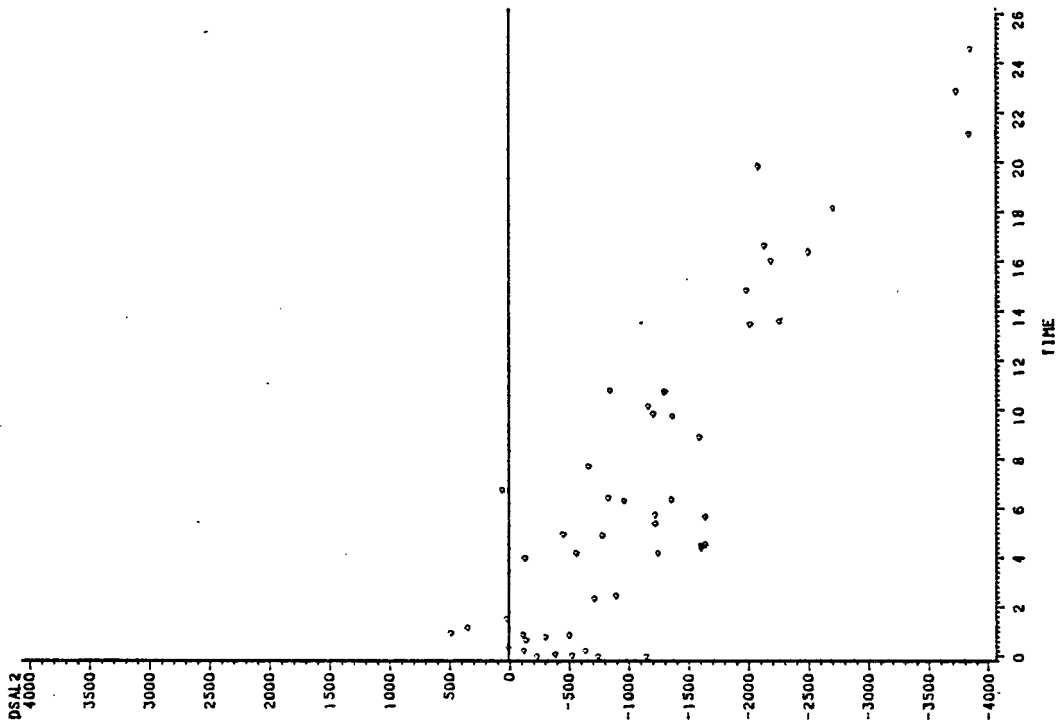
FIGURE 3--FEMALE DSAL PLOTS WITH OVERCORRECTION FOR DISCRIMINATION
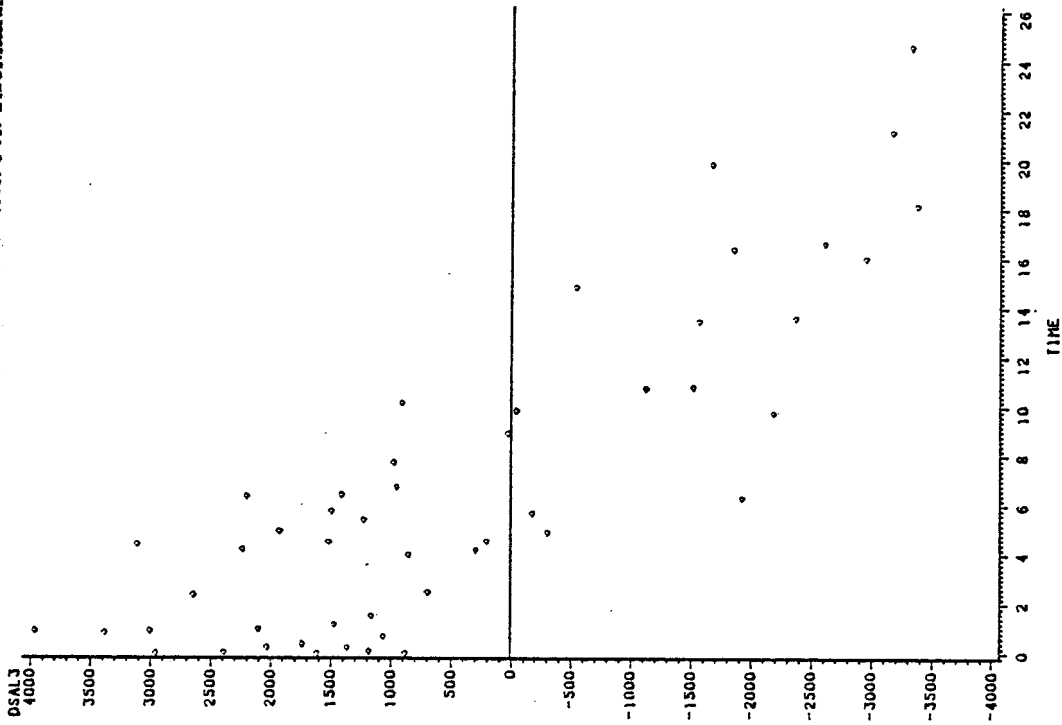
FIGURE 4---FEMALE MEANS PLOTS WITH NO EVIDENCE OF DISCRIMINATION
DSAL MEANS FOR 5 YEAR CLASS INTERVALS

25
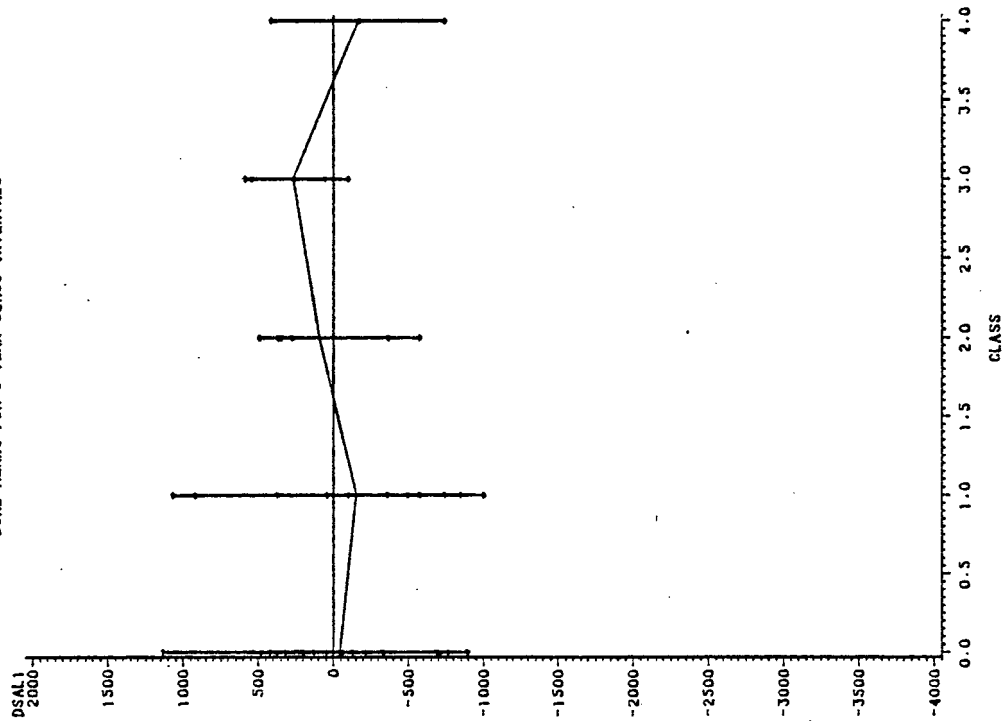
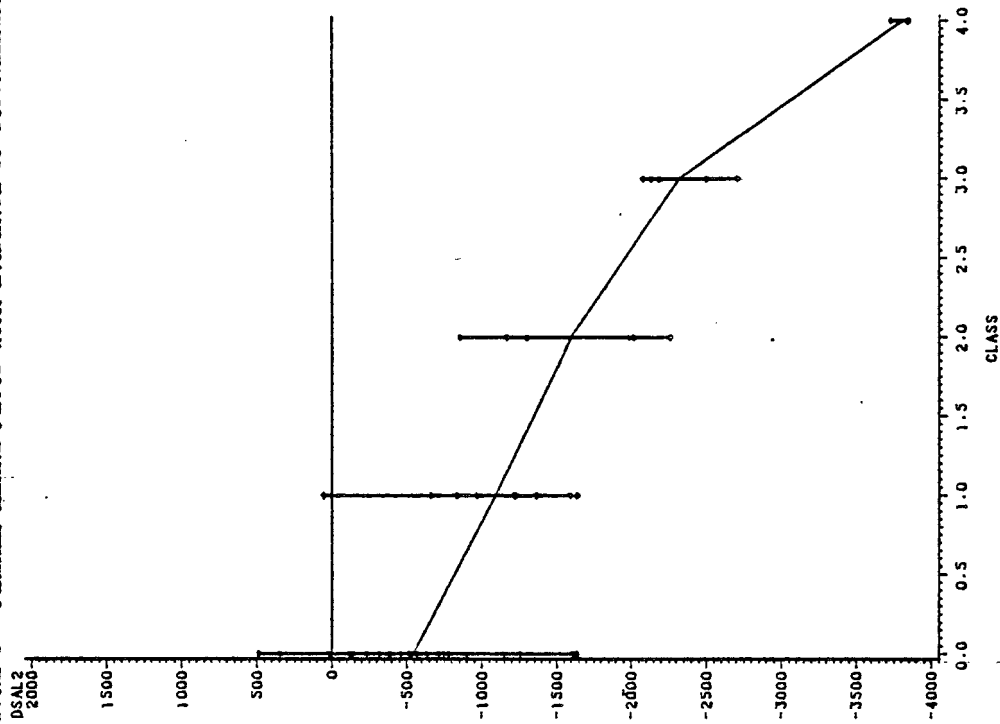FIGURE 5---FEMALE MEANS PLOTS WITH EVIDENCE OF DISCRIMINATION

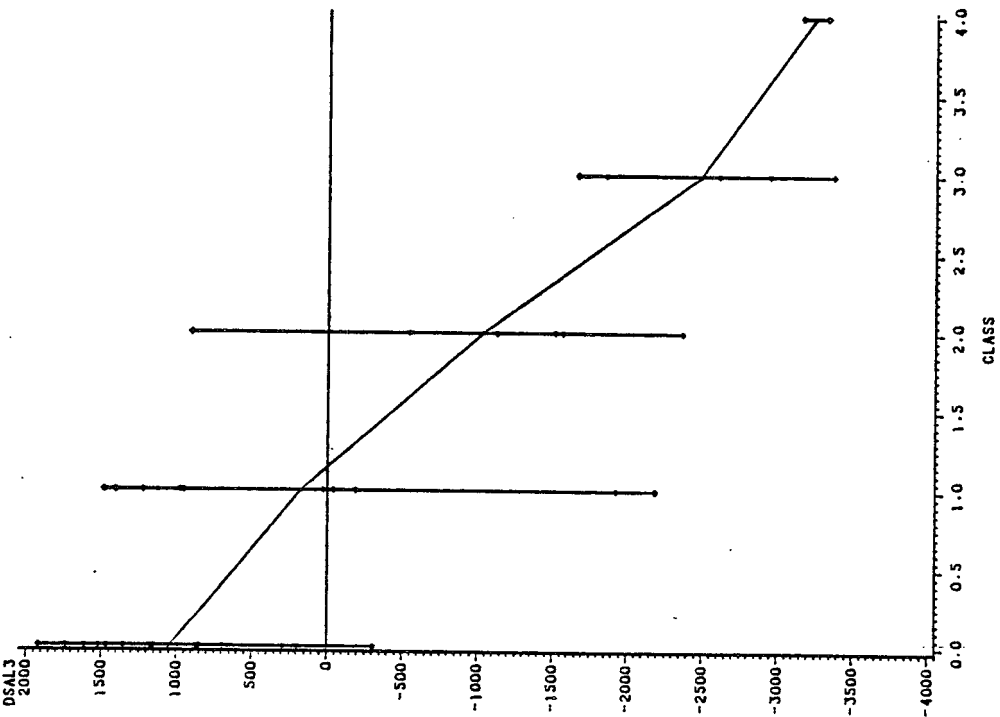FIGURE 6---FEMALE MEANS PLOTS WITH OVERCORRECTION FOR DISCRIMINATION
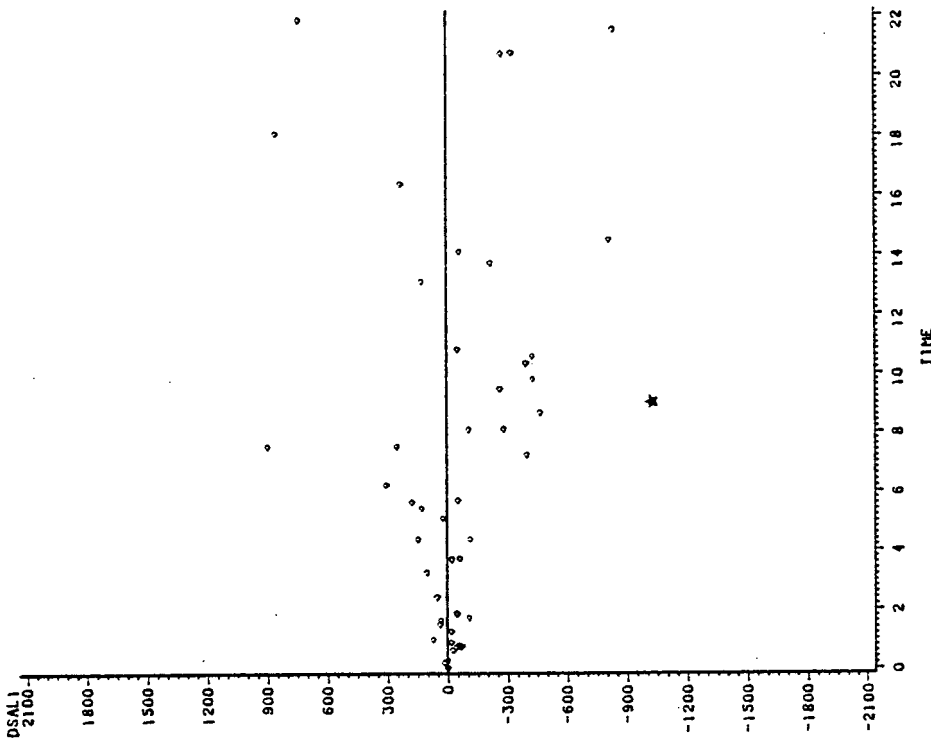
FIGURE 8--PLOT OF FEMALE DSALS FOR NONDISCRIMINATION CASE

FIGURE 9--PLOT OF MALE RESIDUALS SALARIES FOR DISCRIMIONATION CASE
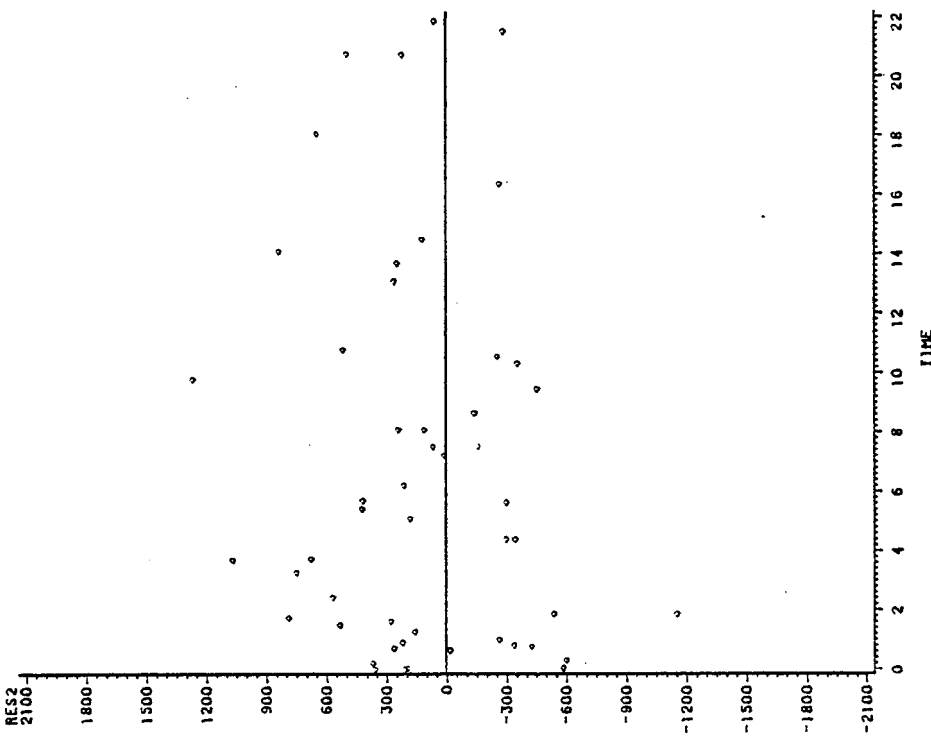
FIGURE 7---PLOT OF MALE RESIDUAL SALARIES FOR NONDISCRIMINATION CASE
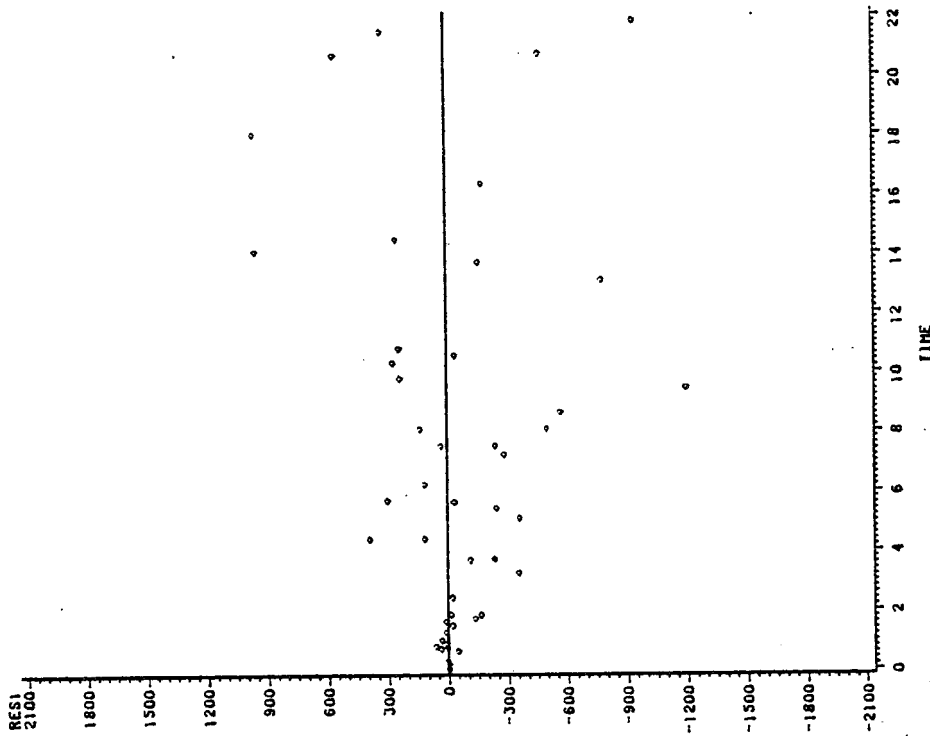
FIGURE 10---PLOT OF FEMALE DSALS FOR DISCRIMINATION CASE
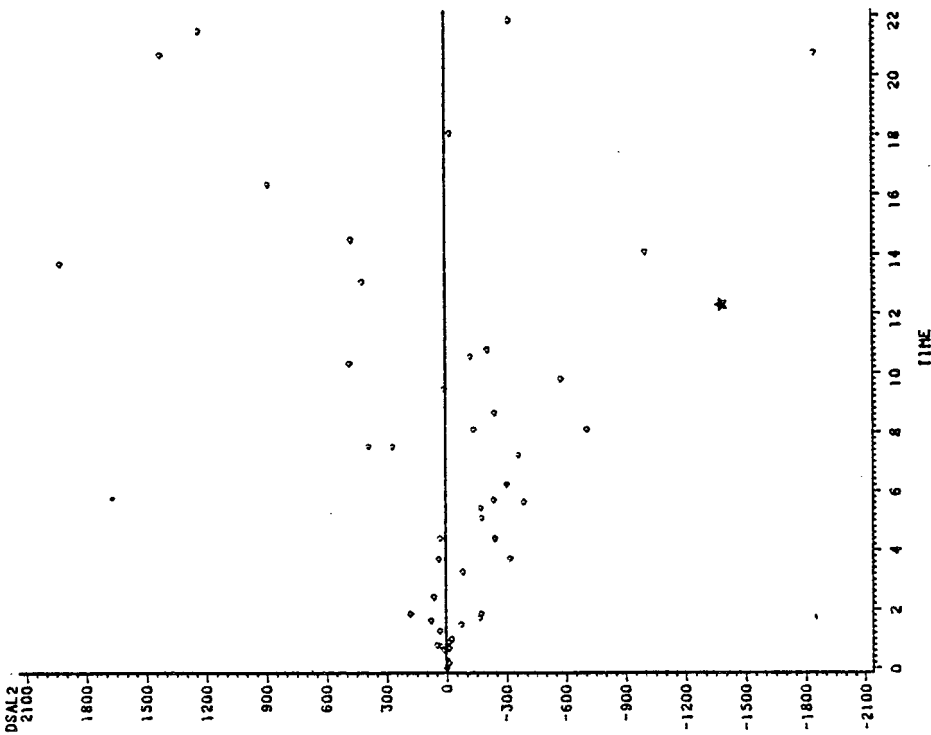
FIGURE 11—TWO SEX MODEL CASE 1

SOLID LINE IS FEMALE MODEL—-DASH LINE IS MALE MODEL
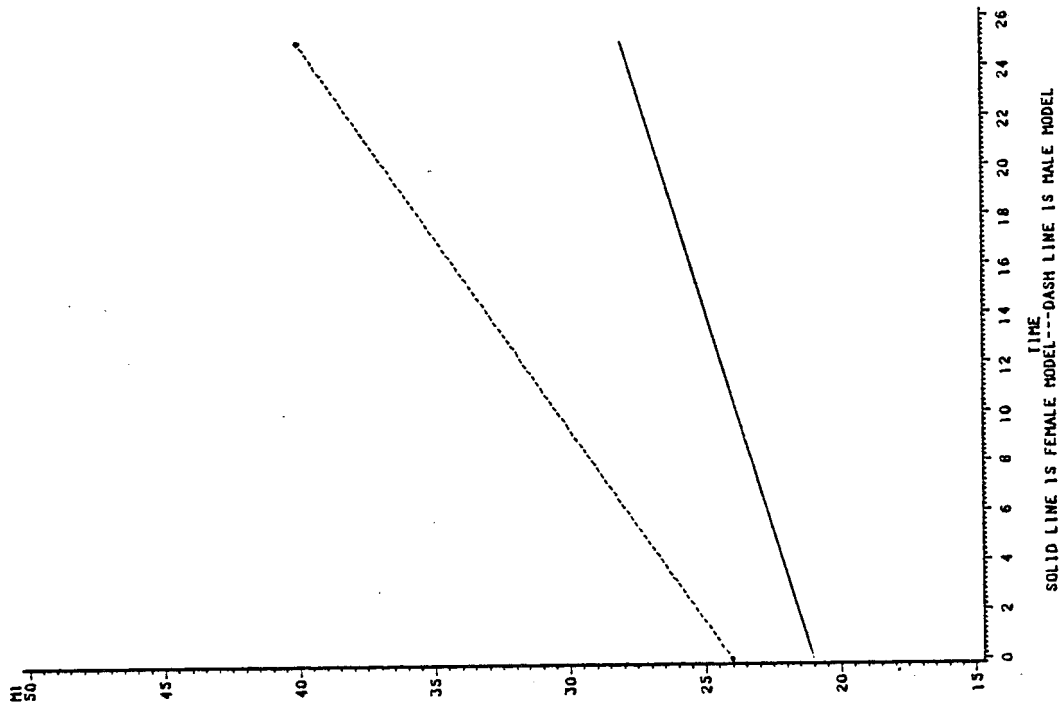
FIGURE 12—TWO SEX MODEL CASE 2

SOLID LINE IS FEMALE MODEL—-DASH LINE IS MALE MODEL
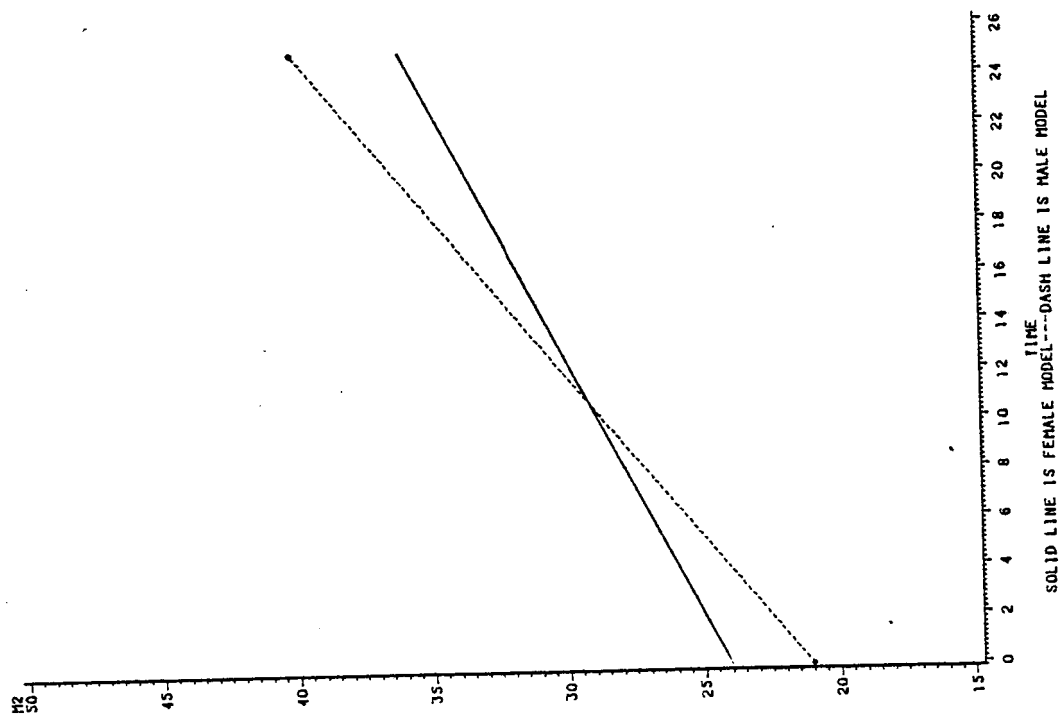
FIGURE 13—TWO SEX MODEL CASE 3

TIME
SOLID LINE IS FEMALE MODEL---DASH LINE IS MALE MODEL
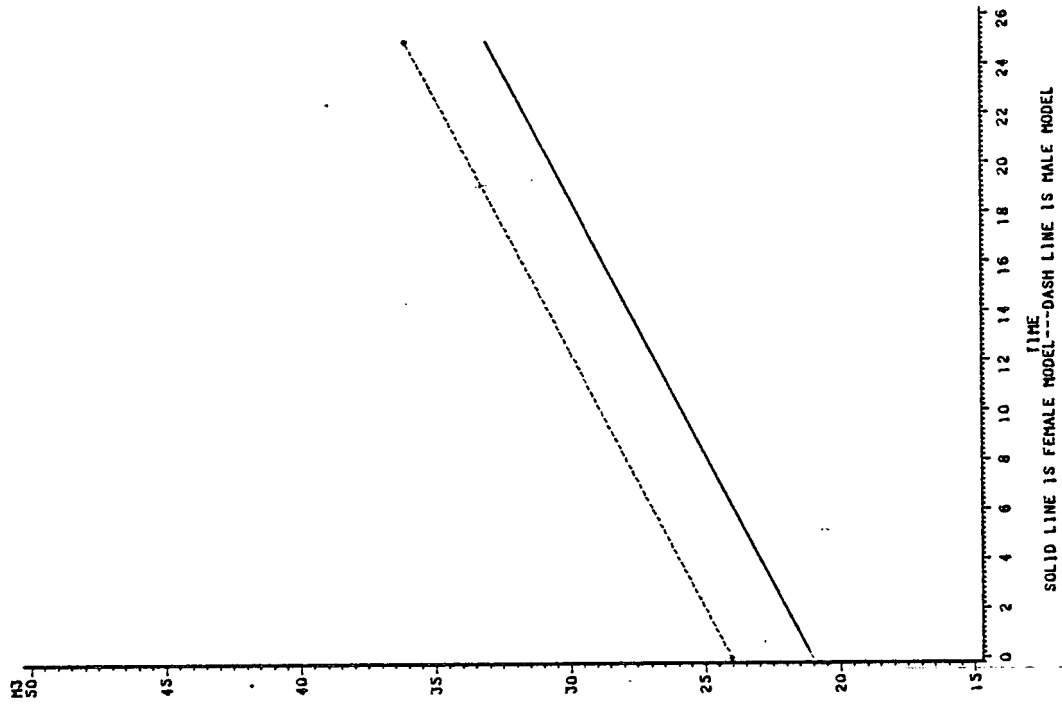
FIGURE 14—TWO SEX MODEL CASE 4

TIME
SOLID LINE IS FEMALE MODEL---DASH LINE IS MALE MODEL
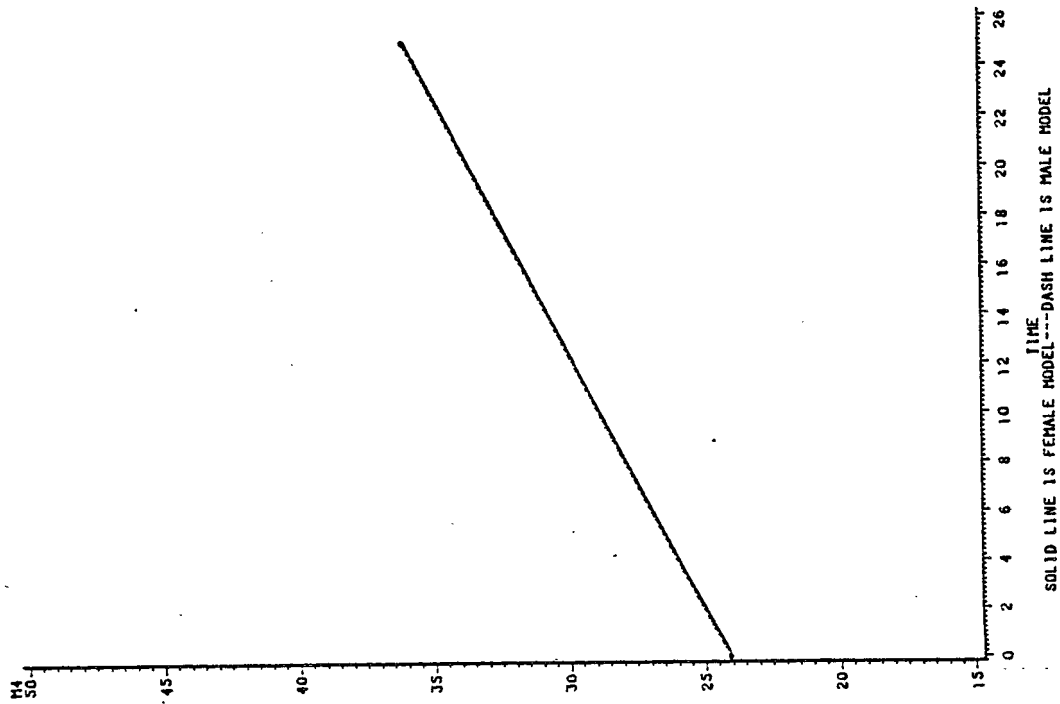
FIGURE 15---FEMALE SALARIES WITH BOTH REGRESSION LINES

SOLID LINE IS FEMALE MODEL---DASH LINE IS MALE MODEL
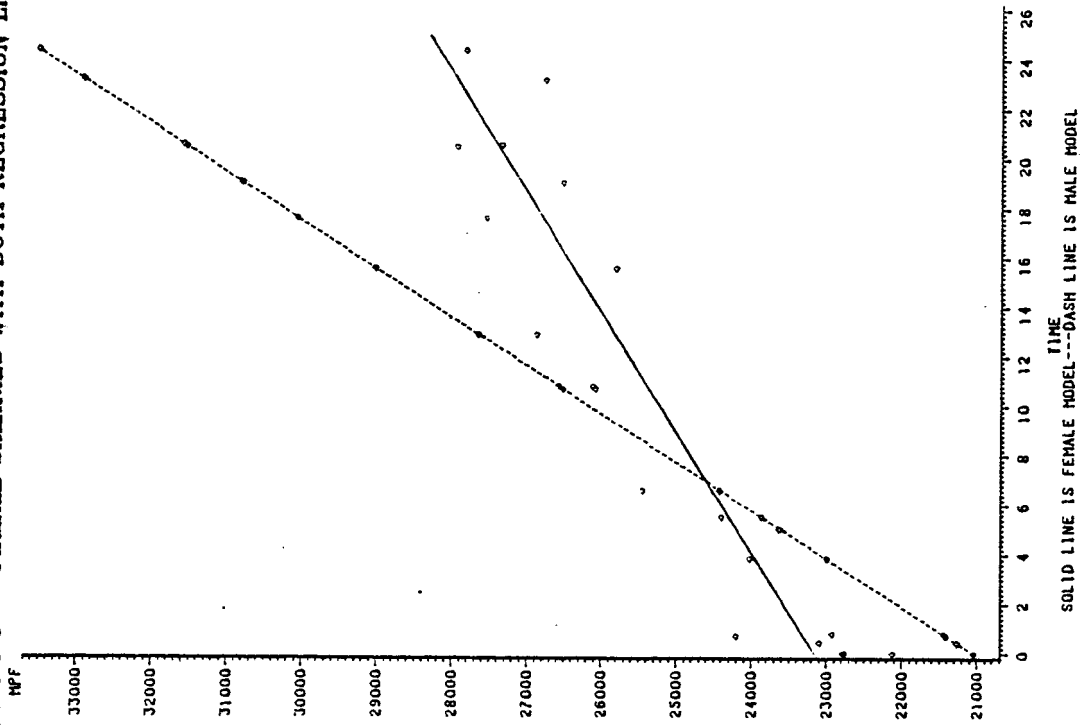
FIGURE 16---FEMALE ADJUSTED SALARIES WITH CONSTANT ADJUSTMENT

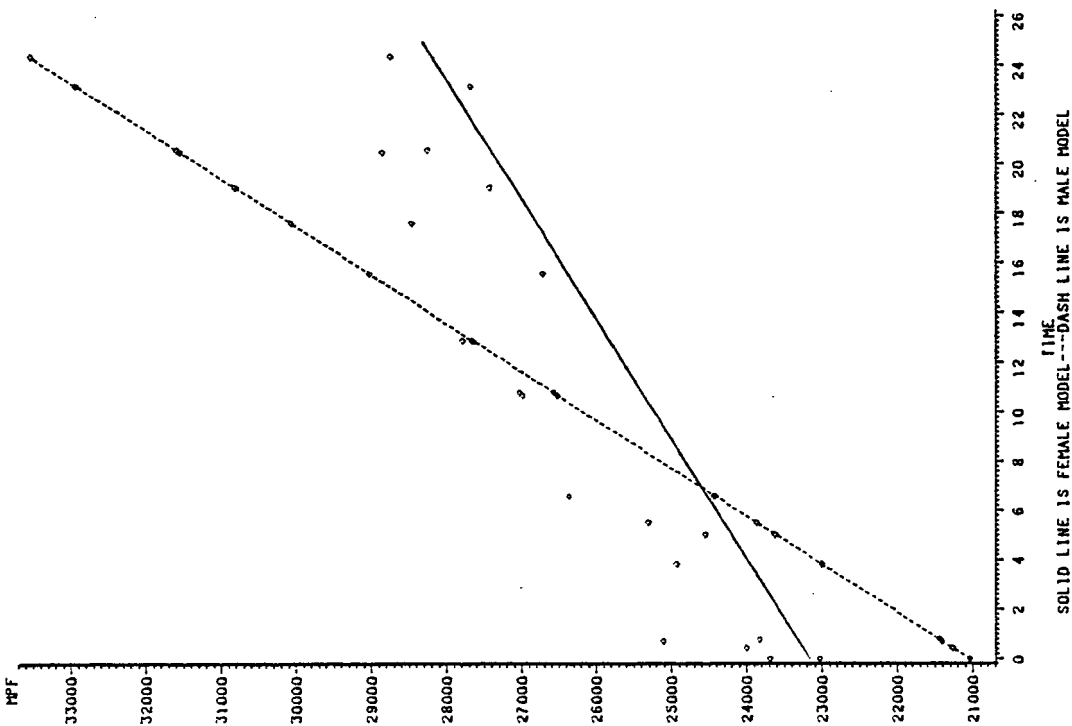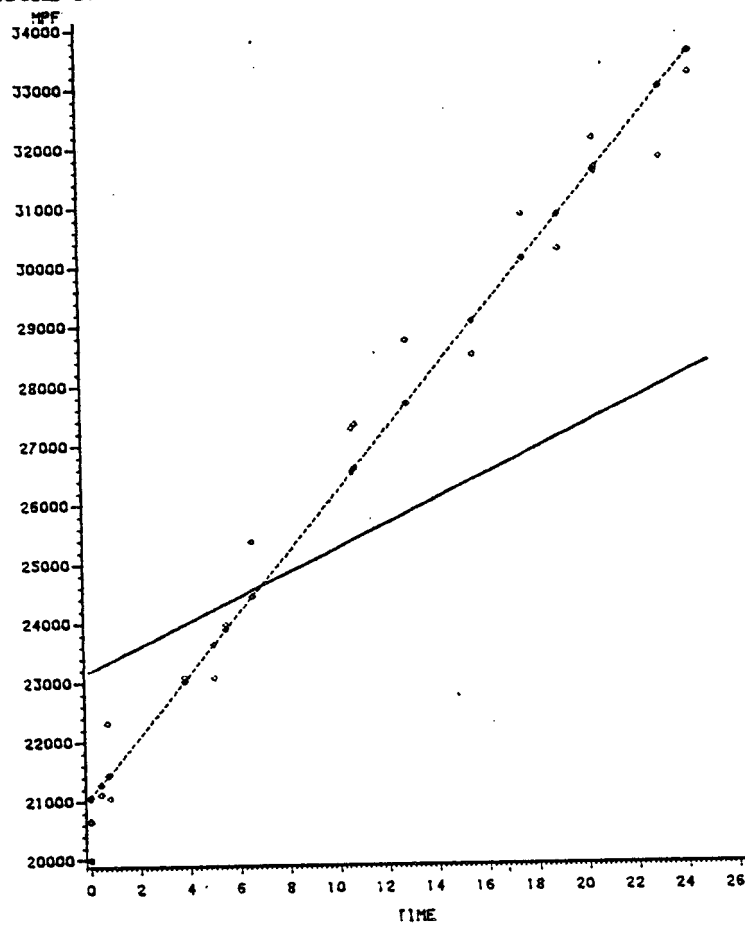SOLID LINE IS FEMALE MODEL---DASH LINE IS MALE MODEL

FIGURE 17--FEMALE ADJUSTED SALARIES WITH EQUATABLE ADJUSTMENT

SOLID LINE IS FEMALE MODEL---DASH LINE IS MALE MODEL

BIBLIOGRAPHY

1. Baldus, D.C., and Cole, J.W.L., (1980). Statistical Proof of Discrimination, McGraw-Hill, New York.

2. Cochran, W.G. (1977). Sampling Techniques, (3rd ed.) Wiley & Sons, Inc., New York.

3. Connolly, W. and Peterson, D., (1980). Use of Statistics in Equal Employment Opportunity Litigation, New York: Law Journal Seminars Press, Inc.

4. Draper, N. and Smith, H., (1981). Applied Regression Analysis (2nd ed.) Wiley & Sons, Inc. New York.

5. Finklestein, M.O., (1979). "The Judical Reception of Multiple Regression Studies in Race & Sex Discrimination Cases." Proceedings of the American Statistical Association: Social Statistics Section.

6. Kaye, D., (1982) "Statistical Evidence of Discrimination." Journal of the American Statistical Association. Vol. 77, No. 380.

7. Kreyszig, E. (1979). Advanced Engineering Mathematics. (4th ed.) Wiley & Sons, Inc. New York.

8. McCabe, G.P., (1979). "The Interpretation of Regression Analysis Results in Sex and Race Discrimination Problems." Proceedings of the American Statistical Association: Social Statistics Section.

9. Reich, D. (1982). "How To Evaluate an Expert's Statistical Analysis." The Practical Lawyer, American Law Institute - American Bar Association, Philadelphia, Vol 28, No. 3.

10. Scott, Elizabeth L., (1979). "Linear Models and the Law: Uses and Misuses in Affirmative Action." Proceedings of the American Statistical Association: Social Statistics Section.

11. Scott, Elizabeth L., (1977). "Higher Education Salary Evaluation Kit." American Association of University Professors, Wash. D.C.